CLUSTERING STUDY OF STR ALLELE DISTRIBUTION PATTERNS

T.W. Wang [1], J. D. Birdwell [1], P. Yadav [1], D.J. Icove [2], S. Niezgoda[3], S. Jones[4]
[1]*The University of Tennessee*
[2]*U.S. TVA Police*
[3]*FBI Laboratory*
[4] *Consultant*

A set of 10,000 DNA STR profiles based on the STR allele probability distribution density functions for Caucasians has been generated for 16-loci. It was assumed that the allele distributions across the loci are independent, and that the occurrence of homozygosity of alleles is also independent of that of all other distribution. The resulting allele distribution has been analyzed using multivariate statistical analysis approach (MVS) to detect clustering patterns among the profiles.

Initial analysis revealed that with the choice of some loci-pair (such as CSFIPO and DL8S51), a clear-cut clustering of the 10,000 profiles emerged. Similarity among the members of each distinct cluster was further studied, to elicit the attributes that made them similar. Differences among the attributes that characterized the clusters were also analyzed to see what set them apart.

Clusterability of DNA profiles based on a priori allele probability density functions has the potential to significantly contribute to the analysis of a target DNA profile in several ways:

- The proper placement of the target profile to within the cluster to which it is most similar, when the characteristics of each cluster are well understood

- The identification of the target profile as an outlier with respect to a set of well established clusters

- The establishment of correlations between the DNA allele distribution patterns and the set of corresponding genetic characteristics associated with the individual possessing the target DNA profile

- The understanding of the evolutionary pathway and divergence of species of animals

Results from the initial clustering study were very encouraging. Among the next steps is the analysis of live data to see if similar clustering patterns hold. In addition, scalability to larger sample sizes while maintaining the clusterability needs to be verified. This work may aid future studies of patterns within the DNA profiles and observed attributes, such as susceptibility to disease or other characteristics, of the corresponding members of the sample population.